Use of Image Subset Features in Image Retrieval with Self-Organizing Maps[*]

Markus Koskela, Jorma Laaksonen, and Erkki Oja

Laboratory of Computer and Information Science, Helsinki University of Technology
P.O.BOX 5400, FI-02015 HUT, FINLAND
{markus.koskela,jorma.laaksonen,erkki.oja}@hut.fi

Abstract. In content-based image retrieval (CBIR), the images in a database are indexed on the basis of low-level statistical features that can be automatically derived from the images. Due to the semantic gap, the performance of CBIR systems often remains quite modest especially on broad image domains. One method for improving the results is to incorporate automatic image classification methods to the CBIR system. The resulting subsets can be indexed separately with features suitable for those particular images or used to limit an image query only to certain promising image subsets. In this paper, a method for supporting different types of image subsets within a generic framework based on multiple parallel Self-Organizing Maps and binary clusterings is presented.

1 Introduction

Content-based image retrieval (CBIR) addresses the problem of finding images relevant to the users' information needs from image databases, based principally on low-level visual features for which automatic extraction methods are available. Due to the semantic gap, i.e. the inherently weak connection between the high-level semantic concepts that humans naturally associate with images and the low-level features that the computer is relying upon, the task of developing this kind of systems is very challenging. One popular method to improve retrieval performance is to use relevance feedback [1], i.e. to adjust the subsequent retrieval process by using information gathered from the user's intra-query feedback.

Another approach to improve retrieval results is to group somehow similar images together and use these groupings to filter out a portion of the non-relevant images for a given query. Unfortunately, in many applications, no semantic annotations or categorizations exist and they are difficult to produce automatically. Still, producing low-level classifications and, in some cases, also certain semantic categorizations are possible with current automatic methods. Examples of low-level classification are distinguishing photographs from computer-generated

[*] This work was supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *New information processing principles*, the latter being part of the Finnish Centre of Excellence Programme.

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 508–516, 2004.

graphics [2,3] and separating color and grayscale images. Certain types of semantic image categories can be distinguished with specialized classifiers which typically perform two-class classifications to the database images. This type of image classification has been studied, for example, to distinguish indoor and outdoor images [4], city images from landscape scenes [5], man-made vs. natural environments [6] and for portraits vs. non-portraits [3]. The effort needed for manual annotation can also be reduced e.g. with active learning [7].

Additionally, such a categorization can be useful in limiting feature extraction to images which are suitable for that particular method. For example, extracting color features may be appropriate only to color images, and shape features requiring segmentation are valid for images containing salient objects and not e.g. for landscape or textural images. A further example is face detection and recognition: in addition to being an important cue of the semantic content in itself, a detected face also makes it viable to extract specific features for face recognition from that image.

In this paper, we extend our existing CBIR system structure to support subset features and binary classifications in addition to the database-wide features used in earlier work. As the main indexing method, we use the Self-Organizing Map (SOM) [8] and propose a uniform framework for incorporating all these feature types into a single system and utilizing them simultaneously and in parallel in image queries. The paper is organized as follows. Section 2 discusses different types of indices needed when some image classification methods and image subsets produced with them are available. In Section 3, the SOM is presented as a common tool for a variety of indices. A set of experiments in which previously recorded user interaction is used as an example case is presented in Section 4. Section 5 then concludes the paper.

2 Indexing Image Subsets

In this section, we consider feature extraction and indexing methods for a whole database and for different types of image subsets. To begin with, it is convenient to identify two fundamental subset types. First, the relevant information of a subset can be contained in set membership, i.e. the subset consists of images having a specific property; or second, the subset contains images for which a certain feature extraction method is either meaningful or available. In the latter case, the mere set membership is semantically insignificant and the pertinent information lies in the internal structure of the subset.

These two types of subsets are clearly not contradictory, but often highly correlated. For example, an object detection module can be used to find images representing a salient object in the foreground and some shape-based features can then be used to describe these images; without prior object detection the feature is useless. On the other hand, whether a photograph is in color or grayscale has often little effect on the semantic content, but e.g. hue-based features give meaningful results only for color images. According to this viewpoint, we can recognize the following three types of image indices.

2.1 Full Indices

Since object and scene recognition and semantic classification in heterogeneous image databases are very difficult problems, the basic features in CBIR are typically extracted from the whole database. Since only a little can be assumed about the image content and the features are to be automatically extracted, such features are usually limited to low-level statistical representations such as global color and texture features.

2.2 Binary Classifications

With some image categorization methods at disposal, we can obtain partitionings of the image database. Commonly, these methods yield two-class classifications for the images; an image either contains or lacks a certain characteristic. A set of different classifications can then be combined to construct a binary index.

An important example of binary classification indices for heterogeneous images is keyword annotation. A straightforward way to utilize keywords describing image contents is to use them as binary attributes affixed to the images. Each keyword divides the database into two subsets: images having and not having that specific keyword in their annotations.

2.3 Image Subsets with Internal Structure

An image subset may also consist of images for which a given feature extraction method is relevant. Extracting these features from all images may be a waste of computational resources or even harmful to retrieval performance. Therefore, such subsets need to be indexed separately. The same holds for situations where a certain feature is available only for a fraction of the database.

One method to gradually improve the retrieval performance of CBIR systems is to utilize information provided by relevance feedback in an inter-query learning scheme. The fact that two images received similar relevance assessments during a specific query is a cue for similarities in their semantic content. With a limited number of recorded query sessions, only a subset of the database has probably been processed in these queries. Still, even a relatively small number of stored queries can improve retrieval results considerably [9]. A similar setting takes place when a portion of the database has been annotated with some accuracy, whereas the remaining images do not contain these annotations.

3 Self-Organizing Map as an Indexing Framework

In indexing methods based on clustering, the data is divided into clusters with the intention that only one or a few of these clusters have to be exhaustively processed in one given query. Typically, each cluster is represented by its centroid or a representative data item and, instead of the original data, the query is first compared to the centroids or cluster representatives. The best cluster or clusters according to the used similarity measure are then selected and the data items belonging to those clusters are evaluated in full. Finally, a fixed number of the most similar items are returned as the result of the query.

This basic clustering approach can readily be extended for our purposes as follows. First, we allow clusterings to be only partial, i.e. there may be data items that do not belong to any cluster, and, second, a data item is allowed to belong to more than one cluster. A set of binary classifications can then be represented with the same data structure, although the individual clusters are now formed of images sharing a certain attribute instead of applying some unsupervised clustering algorithm acting on automatically extracted low-level features.

The Self-Organizing Map (SOM) [8] can also be considered as a clustering method due to the mapping of feature vectors and their associated images to the SOM units. This, however, ignores the topology of the SOM, so a portion of the provided data organization is dismissed. In fact, the distinct strength of the SOM as an indexing method lies in its property of topology preservation. The SOM preserves topology on the map grid and this enables the spreading of the user-provided relevance assessments also to the neighboring map units since they can be assumed to contain similar feature vectors and thus similar images.

3.1 PicSOM

The PicSOM [10,11] image retrieval system is a framework for research on content-based image retrieval. As the name implies, PicSOM uses the SOM as its basic image indexing method, although other clustering methods are also supported. For example, k-means clustering was experimented with and compared to the SOM in [11]. Instead of the standard SOM version, PicSOM uses a special form of the algorithm, the Tree Structured Self-Organizing Map (TS-SOM) [12]. The hierarchical structure of TS-SOM is useful for two purposes. First, it drastically reduces the complexity of training large SOMs needed for indexing large databases by exploiting the hierarchy in finding the best-matching map unit (BMU) for an input vector. Second, the hierarchical representation of the image database produced by a TS-SOM can be utilized in browsing and visualizing the images in the database.

3.2 Multiple Self-Organizing Maps

The PicSOM system is fundamentally based on using several parallel SOMs trained with separate feature data simultaneously in image retrieval. The features are usually comprised of statistical visual data such as the MPEG-7 [13] content descriptors. Any additional vectorial data can, however, be used to train corresponding SOMs and thus be used in image retrieval. If the feature in question was extracted only from a subset of the images in the database, only that subset is used in the training the corresponding SOM. The size of the SOM should be set accordingly, as it is not reasonable to use SOMs of the same size with small subsets as with the whole database.

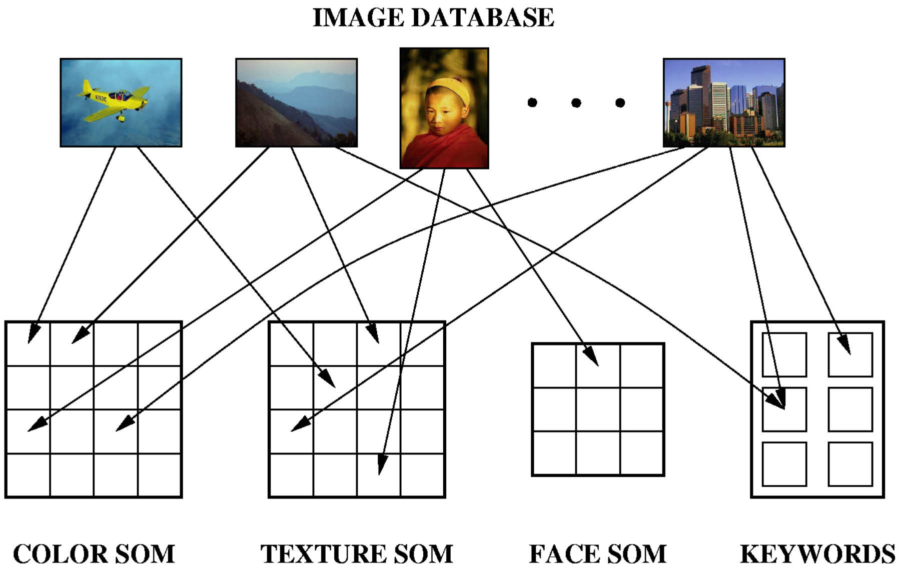

Fig. 1. An example of using four parallel indices for an image database. The color and texture SOMs are trained with the whole database and the face SOM with a subset obtained with face detection. Keyword annotations are also available for some images.

After training the SOMs, their map units are connected with the images of the database. This is done by locating the BMU for each image on each SOM. As a result, the different SOMs impose different similarity relations on the images and the system thus inherently uses multiple features for image retrieval. An illustration with two full-database SOMs, one SOM trained with an image subset, and keyword-based binary classifications is presented in Figure 1.

3.3 Relevance Feedback with Multiple Feature Indices

The relevance feedback mechanism of PicSOM is a crucial element of the retrieval engine. The basic method is only briefly presented here, see [10] for a comprehensive treatment. During a retrieval session, the user marks images that she considers relevant as positive, and the remaining ones are implicitly regarded as negative. As the first step, the SOM units are awarded a positive score for every relevant image mapped in them resulting in an attached positive impulse. Likewise, associated non-relevant images result in negative scores and impulses. If the total numbers of relevant and non-relevant shown images are $N^+(n,m)$ and $N^-(n,m)$ at query round n on mth SOM, the positive and negative scores are simply the inverses: $x_+(n,m) = 1/N^+(n,m)$ and $x_-(n,m) = -1/N^-(n,m)$. For each SOM, these values are mapped from the shown images (and thus rated either as positive or negative) to their corresponding BMUs where they are then summed. This way, we obtain a zero-sum sparse value field on every SOM in use. With SOMs trained with image subsets, we neglect the shown images that are

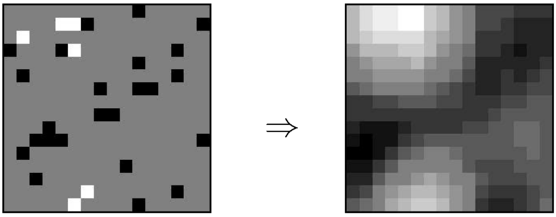

Fig. 2. An example of how a SOM surface is convolved with a tapered window function. On the left, images selected and rejected by the user are shown with white and black marks, respectively. On the right, the convolution result, where relevance information is spread around the centers.

not mapped to that particular SOM. Since the sparse value fields are zero-sum, we introduce no bias against the non-indexed images.

Due to the topology preservation of the SOM, we are motivated to spread the relevance information (both positive and negative) provided by the user also to the neighboring map units of the BMUs. This can be done by convolving the sparse value fields with tapered (e.g. triangular or gaussian) window functions. Figure 2 illustrates how the positive and negative responses are first mapped on a 16×16-sized SOM and how these responses are expanded in the convolution.

The indices formed by binary classifications are treated similarly with two exceptions. First, since no topology exists, the spreading of user responses with convolution is not valid (or, we always use unit impulse as the convolution window). Second, the same image may be present in multiple clusters, which is taken into account by dividing the relevance weight equally to all these binary classes.

As the response values of the parallel indices are mutually comparable, we can determine a global ordering for determining the overall best candidate images. By locating the corresponding images in all indices, we get their scores with respect to different feature extraction methods. The total qualification values for the candidate images are then obtained simply by summing the corresponding responses. Content descriptors that fail to coincide with the user's conceptions mix positive and negative user responses in the same or nearby map units and binary classes. Therefore, they produce lower qualification values than those descriptors that match the user's expectations and impression of image similarity and thus produce areas or clusters of high positive response. As a consequence, the parallel content descriptors and indices do not need explicit weighting.

4 Case Study: Recorded User Interaction

In the following experiments, we use a database of $N = 59\,995$ miscellaneous images from Corel Photo CDs. We created manually the following six image classes as ground-truth: **faces** (1115 images, *a priori* probability 1.85%), **cars** (864 images, 1.44%), **planes** (292 images, 0.49%), **sunsets**, (663 images, 1.11%), **horses**, (486 images, 0.81%), and **traffic signs**, (123 images, 0.21%). As visual features, we used a subset of MPEG-7 [13] descriptors, viz. *Scalable Color*, *Dominant Color*, *Color Structure*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture*, and *Region Shape*. These descriptors were extracted from every image in the database, and 256×256-sized SOMs were then trained for each of them.

As an example case, we study the utilization of previously recorded relevance evaluations provided by users during earlier query sessions (inter-query learning). The relevance evaluations provided by the user during a query session partition the set of shown images into relevant and nonrelevant classes with respect to the target of that particular query. Specifically, when two images both have been marked as relevant within the same query, we can assume that the semantic contents of the images are somehow similar. In the experiments, the relevant images of individual queries are used as image clusters. The relevance evaluations consisted of 317 saved query sessions in which a total of 6897 images (11.5% of the database) had been marked relevant at least once. On the average, a recorded query contained 25.8 relevant images. As the first one of the studied methods, we use these image clusters directly as a binary classification index.

Alternatively, we use the evaluations as statistical features by constructing a fixed-length binary feature vector for images with recorded evaluations so that each dimension of the vector corresponds to a recorded query. Since the remaining images do not have any stored assessments, they are omitted from this inter-query index. This way, we obtain 317-dimensional inter-query feature vectors for the 6897 images. Thirdly, as a preprocessing step, we reduce the dimensionality of these feature vectors to 50 with singular value decomposition (SVD). For comparison, we train two 64×64-sized SOMs for the feature vectors, one before and the other after the dimensionality reduction.

In the test setting, each image in the studied six ground-truth classes was used one at a time as an initial reference image for category search. The system returned 20 images at each round, and with 50 rounds per query session the total number of shown images was $N_T = 1000$ images, i.e. 1.67% of the database size. Relevance feedback was used to refine the query as categorical feedback, i.e. images belonging to the studied class were marked as relevant and others as nonrelevant. This way, the retrieval experiments could be carried out automatically. In spreading the responses of the sparse value fields, triangular windows of 6 and 8 map units in length were used for the user interaction SOMs and the other SOMs, respectively.

The averaged recall–precision plots for the six ground truth classes are shown in Figure 3. The MPEG-7 descriptors are used in all cases, either solely or with one of the three inter-query indices. Overall, it can be seen that due to the semantic gap, the precision of using the low-level features alone remains modest, especially since no segmentation was used. Using information provided by the recorded queries considerably improves retrieval precision regardless of the used method, even though only 11.5% of the images are included at all in the recordings. Using the recorded queries as binary classifications yielded better results than their use as statistical feature vectors with SOMs. This, however, comes with an increase in computational requirements, since the shown images are often mapped to multiple clusters. Also, the approach does not scale well to

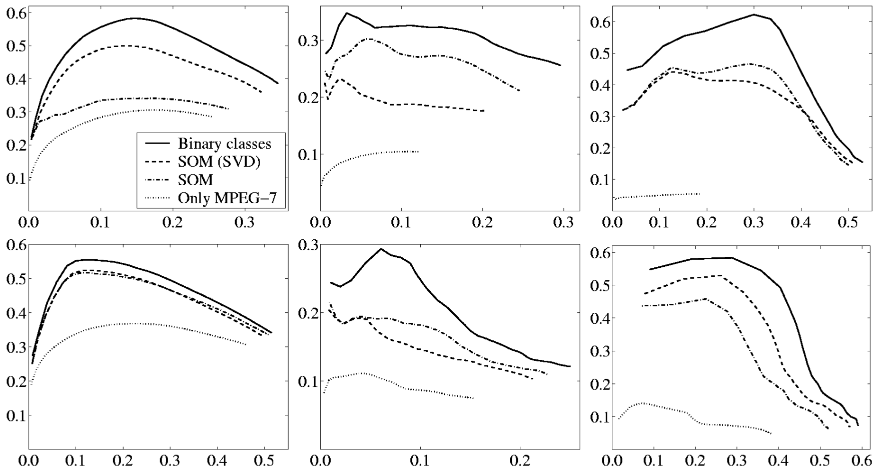

Fig. 3. Recall–precision plots (x-axis: recall; y-axis: precision) using the MPEG-7 descriptors solely and with recorded queries as binary classifications and as subset features with and without dimensionality reduction (SVD). Used classes were (top row, left-to-right) **faces**, **cars**, **planes**, (bottom row) **sunsets**, **horses**, and **traffic signs**.

large amounts of recorded data since the number of binary classification clusters equals the number of recorded queries.

The use of a subset feature SOM based on the recorded queries also improves the results significantly. With this approach, the increase in online computational requirements is minor since we only add one SOM index along the seven parallel SOMs trained with the MPEG-7 descriptors. Reducing the dimensionality of the inter-query feature by SVD before the SOM training does not systematically alter the retrieval results; with **faces** and **traffic signs** the dimensionality reduction improves results, with **cars**, **planes**, and **horses** the results are worse. In an application where a lot of recorded usage data could easily be accumulated, this kind of preliminary dimensionality reduction would be essential as the dimensionality of the training data equals the number of recorded image queries, which could well be in the order of thousands or more.

5 Conclusions and Future Work

With large databases of general images, the retrieval performance of low-level visual features alone often remains quite modest (as is observed also in Fig. 3) and additional feature types can be highly beneficial. One method for improving results is to incorporate some automatic image classification methods to the retrieval system. The resulting subsets can then be indexed separately or used to limit the query only to specific image subsets.

In the experiments of this paper, previously recorded query sessions were interpreted either as binary classifications or as statistical features for an image subset and used in parallel with visual MPEG-7 descriptors. The experiments showed that both approaches produced clearly improved results and that using this information greatly enhances the precision of the system without any additional burden to the user. While using previously stored retrieval sessions performed by assorted users of the system might conflict with the subjectivity and context-dependency of human notion of image similarity, in practice the user assessments provide valuable accumulated information about image semantics.

A straightforward direction for future work is to include automatic semantic classifications to the PicSOM system and study whether indexing these subsets separately would be beneficial. For this purpose, a generic framework of feature indices, both database-wide and subset-based, was presented in this paper.

References

1. Zhou, X.S., Huang, T.S.: Relevance feedback for image retrieval: A comprehensive review. Multimedia Systems **8** (2003) 536–544
2. Frankel, C., Swain, M.J., Athitsos, V.: Webseer: An image search engine for the world wide web. Technical Report 96-14, The University of Chicago (1996)
3. Gevers, T., Aldershoff, F., Geusebroek, J.M.: Integrating visual and textual cues for image classification. In: Proceedings of Fourth International Conference on Visual Information Systems (VISual 2000), Lyon, France (2000) 419–429
4. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, Bombay, India (1998) 42–51
5. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1935
6. Yoon, J., Jayant, N.: Semantics-sensitive image retrieval: An information fusion approach. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2003). Volume 1., Baltimore, MD, USA (2003) 761–764
7. Sychay, G., Chang, E., Goh, K.: Effective image annotation via active learning. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2002). Volume 1., Lausanne, Switzerland (2002) 209–212
8. Kohonen, T.: Self-Organizing Maps. Third edn. Springer-Verlag (2001)
9. Koskela, M., Laaksonen, J.: Using long-term learning to improve efficiency of content-based image retrieval. In: Proc. of Third International Workshop on Pattern Recognition in Information Systems (PRIS 2003), Angers, France (2003) 72–79
10. Laaksonen, J., Koskela, M., Laakso, S., Oja, E.: Self-organizing maps as a relevance feedback technique in content-based image retrieval. Pattern Analysis & Applications **4** (2001) 140–152
11. Laaksonen, J., Koskela, M., Oja, E.: PicSOM—Self-organizing image retrieval with MPEG-7 content descriptions. IEEE Transactions on Neural Networks, Special Issue on Intelligent Multimedia Processing **13** (2002) 841–853
12. Koikkalainen, P., Oja, E.: Self-organizing hierarchical feature maps. In: Proceedings of International Joint Conference on Neural Networks. Volume II., San Diego, CA, USA (1990) 279–284
13. MPEG: MPEG-7 Overview vers. 9 (2003) ISO/IEC JTC1/SC29/WG11 N5525.